NCBI Molecular Biology Resources

South Central Region NNLM

September 15, 2010

Topics for Today

- About NCBI
- Overview of Molecular Databases
  - OMIM, GenBank, RefSeq, Gene, SNP, Structure
- Apolipoprotein E example
  - Using the Entrez system
  - Using BLAST

The National Center for Biotechnology Information

Bethesda,MD

Created in 1988 as a part of the National Library of Medicine at NIH

- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biomedical information

Aspects of Molecular Data

- Sequences
- Expression
- Genome Maps
- 3D Structures
- Protein Domains
- Homologous Genes, Proteins, Structures
- Pathways
- Genetic Variation

NCBI Public Services

Selected NCBI Molecular Databases

- Biomedical literature
  - PubMed free Medline
  - PubMed Central full text online access
  - OMIM human disease genes
  - NCBI Bookshelf online biomedical textbooks
- Biomolecular Databases
  - GenBank largest sequence database
  - RefSeq curated NCBI reference sequences
  - dbSNP small scale genetic variations
  - MMDB NCBI's 3D structure database
  - GEO microarray expression data

NCBI Public Services

Metadatabases: Information Hubs

- UniGene sequence based gene catalog
- Taxonomy master browser for molecular data
- Gene molecular and literature related to genes

NCBI Public Services

Selected NCBI Search Services and Tools

- Entrez integrated literature and molecular databases
- BLAST highest volume sequence search service
- VAST structure similarity searches
- Map Viewer graphical genome map display

Global Entrez Query: All NCBI Databases

The Entrez system: 40 (and counting) integrated databases

Entrez: A Discovery System

Word weight

PubMed abstracts

VAST
Neighbors
Related Structures

Phylogeny

Pre-computed and pre-compiled data.
- A potential "gold mine" of undiscovered relationships.
- Used less than expected.

BLAST
Neighbors
Related Sequences

Nucleotide sequences

Protein sequences

Hard Link

BLAST
Neighbors
Related Sequences
BLink
Domains

3

The Discovery Initiative

- Easier to use interfaces
- Promote higher quality resources
  - Gene
  - RefSeqs
- Expose the power of pre-computed similarities and pre-compiled links

---

Discovery Components in Entrez

- **Database Ads** – direct to related information in other database
- **Sensors** – point to other databases or special search tools where the query is more relevant
- **Analysis tools** – access to live analysis results

Apolipoprotein E (APOE)

- Important serum lipid transport protein.
- Defects implicated in cardiovascular disease and late-onset. Alzheimer disease (LOAD).
- Three common isoforms (alleles).

Isoform	Position 112 (130)	Position 158 (176)
e3	Cysteine (C)	Arginine (R)
e4	Arginine (R)	Arginine (R)
e2	Cysteine (C)	Cysteine (C)

The e4 isoform (allele) is associated with increased risk of LOAD

Human APOE Data Goals

- Reference Sequences
  - transcript (mRNA)
  - protein
  - gene (gDNA)
- Sites of Expression
  - UniGene Expression profile
  - GEO profiles
- View genomic assemblies and Maps
  - Comparative Maps

- Disease gene polymorphisms
- Genotypes
  - Reference Genome
  - HuRef (JC Venter genome)
- Homologs in other species
  - HomoloGene
  - BLAST to find Panda homologs

Live demonstrations

www.ncbi.nlm.nih.gov

Human APOE Gene and RefSeqs

1. Search All Databases with APOE from NCBI Homepage
2. Retrieve PubMed results
3. Follow Gene Ad to Gene ID 348
4. Follow link to "reference sequence details"

- Transcript and protein
  - NM_000041.2
  - NP_000032.1
- Genomic
  - RefSeq Gene NG_007084.2
- Genome Builds
  - Reference NC_000019.9
  - HuRef

Genome Maps

APO Gene Cluster

1. Gene links menu to Map Viewer
2. Maps and Options
   1. Remove all but Gene
3. Zoom out 10X using graphic
4. Zoom our 2X by clicking on map

Comparative Maps

1. Map Viewer Maps and Options
   1. Add Chimp and Mouse Gene Maps

Allelic Variants and Disease

OMIM

- Gene links menu to OMIM
- Retrieve APOLIPOPROTEIN E; APOE
- Click allelic variants
- Jump to .0016 ALZHEIMER DISEASE 2 [APOE, CYS112ARG]
- Recent activity back to APOE gene record

dbSNP

- Gene links menu to SNP: Gene View
- Check "Include clinically associated" and click "Refresh"
- Position 130 Cys->Arg
- Click on rs429358 in table
- Note HuRef has C allele coding for Arg

Expression Information

UniGene

1. Gene links menu to UniGene
2. EST profile
3. Widespread expression

GEO Profiles

1. Gene links menu to GEO profiles
2. Use History to combine link from Gene AND GDS596
3. Predominate expression in liver, some in brain

Homlogous Genes / Proteins

HomoloGene
1. Gene links menu to HomoloGene
2. Use "pairwise alignments" device to compare Human and Chimp proteins at position 130
3. Use UniGene portion to find rabbit (*Oryctolagus cuniculus*) homolog

Using BLAST
1. Gene link to RefSeq protein NP_000032
2. "Run BLAST" from Discovery Column
3. Choose Reference Proteins database
4. Organism limit Giant Panda
5. BLAST button

Finding a Structure

Related Structures
1. Gene link to NP_000032
2. Links menu to Related Structures
3. Change to "All similar MMDB" and click "Go"
4. Link to sequence alignment from 1B68_A
5. Position 112 polymorphism

Structure and Cn3D
1. Related structures link to structure for 1B68
2. Click Structure View in Cn3D (Cn3D must be installed)
3. Manipulate structure
4. Add side chains using Style -> Edit Global Style
5. Highlight Arg 112

Service Addresses

- *General Help*   `info@ncbi.nlm.nih.gov`
- *BLAST*   `blast-help@ncbi.nlm.nih.gov`

Telephone support: 301- 496- 2475